DOCUMENT RESUME

ED 198 177 TM 810 145

AUTHOR Leinhardt, Gaea; Seewald, Andrea Mar
TITLE Overlap: What's Tested, What's Taught?
INSTITUTION Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY National Inst. of Education (ED), Washington, D.C.
REPORT NO LRDC-1980/16
PUB DATE Jun 80
NOTE 32p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Curriculum Evaluation; *Evaluation Methods; *Instruction; *Program Effectiveness; Teacher Attitudes; *Testing Problems
IDENTIFIERS Test Curriculum Overlap

ABSTRACT

In studying the effectiveness of different instructional practices or programs, it is possible that an outcome measure may be biased in favor of a particular practice or program because the overlap between the test and one program is greater than for the other(s). Two basic approaches for dealing with the overlap issue are reviewed. The first approach is a systematic analysis of curricula and tests to help guide test selection. A second approach to overlap is to measure the degree of overlap directly, and incorporate such a measure into the analysis. This approach could be used in conjunction with the first, or alone. Also reviewed are two ways to directly measure overlap which have been developed. The first involves teacher interviews or questionnaires. The second involves analyzing the curriculum to assess if information required by the test has been covered by the curriculum. The teacher interview approach reflects both informal in-class instruction and curriculum-based instruction, but it may also include the teacher's expectation about student competency. Both approaches do equally well in predicting final test performance. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED198177

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

LRDC

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

LEARNING RESEARCH AND DEVELOPMENT CENTER

OVERLAP: WHAT'S TESTED, WHAT'S TAUGHT?

1980/16

GAEA LEINHARDT AND ANDREA MAR SEEWALD

University of Pittsburgh

OVERLAP: WHAT'S TESTED, WHAT'S TAUGHT?

Gaea Leinhardt and Andrea Mar Seewald

Learning Research and Development Center

University of Pittsburgh

June, 1980

The research reported herein was supported by the Learning Research and Development Center, supported in part by funds from the National Institute of Education (NIE), United States Department of Health, Education, and Welfare. The opinions expressed do not necessarily reflect the position or policy of NIE, and no official endorsement should be inferred.

FEB 2 3 1981

Abstract

In studying the effectiveness of different instructional programs, it is possible for a criterion measure to favor one program over the others because the overlap between the criterion test and program content is greater. If the overlap is not controlled for, one program may look artificially better with respect to test performance than others. Several approaches to dealing with overlap are reviewed and two techniques for its estimation, in the context of actual educational evaluations, are explored.

OVERLAP: WHAT'S TESTED, WHAT'S TAUGHT?

Gaea Leinhardt and Andrea Mar Seewald

Learning Research and Development Center

University of Pittsburgh

Considering the extent to which school children are tested in the United States, and the myriad decisions that are based on the results of such tests, it is important to understand what generates variations in test performance. Decisions range from those that affect the individual student to those that affect a school or district. Examples of individual decisions include: passing a student on to new material, promotion to the next grade, and eligibility for placement in special classes. Examples of district level decisions include: choosing a new curriculum, selecting schools for special services, or adopting a new compensatory program. Because testing information is the basis for individual and program evaluation, understanding the real meaning of a test score is of considerable importance.

The score received on a "good" achievement test is meant to reflect the actual knowledge that an individual or group has about the domain from which that test was drawn. That is, a good test is assumed to be a random sample of items from a specified domain, where domain includes both content covered and item form. The domain of instruction may be identical to the domain sampled by the test, it may

partially overlap or share the domain of the test, or it may be totally different. When a set of test scores are used to help evaluate the impact of instructional programs, knowledge about the extent of overlap is critical to interpretation of the results. If different instructional programs have different overlap with the criterion measures then results can be biased in favor of the program with the higher overlap. The purpose of this paper is to review approaches to the problem of dealing with overlap, to suggest methods for measuring overlap between test and curriculum, and to examine the implications of overlap for evaluation studies.

Approaches to the Problem of Test-Curriculum Overlap

Educators, especially curriculum designers, have long been aware that tests and curricula can and often do emphasize different aspects of a particular knowledge domain (Cole & Nitko, 1979; Rosenshine, 1978; Walker & Schaffarzick, 1974). Awareness of the problem of the fit or lack of it between curricula and tests has generated four types of solutions: first, to build new "criterion-based" tests for each situation; second, to alter existing tests by deleting items that do not reflect curriculum content; third, to systematically analyze curricula and tests in order to select the best existing test; and fourth, to directly measure the relationship and incorporate it into an analysis.

Build Tests to Match Curricula

Popham (1978) has been the main proponent for the construction of "new" criterion-referenced tests to accurately reflect the content of different curricula. This approach is advantageous because it insures maximum fit. However, program evaluations often involve multiple contrasts. When this is the case, the approach of building specific tests must be modified to produce a single test that is in some sense comparable or fair to all curricula. For example, if four programs were to be contrasted, each of which used a different curriculum, a battery could be constructed by selecting items unique to each curriculum, shared by each curriculum, and untaught by any curriculum. Of course, in order to be fair, the test should be equally weighted for each curriculum in terms of the proportion of items of each type. Such a test would assure that some predetermined quantity of the test had been covered and that some proportion of the content covered was actually tested. Kugle and Calkins (1976) describe a procedure for developing criterion referenced tests for fifth grade math and social studies classrooms by matching subtests and curricular objectives. Test construction seemed difficult and costly and would be so in most cases. Also, in the work of Kugle and Calkins, only one instructional program was being used. Most program interventions use several curricula in different configurations which further complicates test construction. If, however, only one or two curricula are being contrasted, building new tests or modifying existing ones may be straightforward and useful. There are, of course, tremendous advantages in moving from norm-referenced to criterion-referenced

standardized tests. The issue here is only whether to build new tests "on the spot" or not.

Alter Existing Tests

Another approach is to take existing test batteries and delete items not included in a particular curriculum or to record separate scores for material taught and not taught. While such an approach may be quite informative for formative evaluations when developers are in a curriculum revision mode, it is less useful for summative work. First, if several curricula are contrasted, then selecting items for inclusion or exclusion is as difficult as constructing new test items. Second, "messing" with a test by removing items dilutes or destroys the desirable psychometric properties of that test. Third, deleting items does not guarantee that what is taught gets tested, only that what is tested was taught.

Selecting the Best-Fitting Test

Formal attempts to deal directly with the overlap problem by identifying the best-fitting test have developed along two rather different lines: detailed curriculum analysis and teacher-based estimations. The curriculum analysis approach has grown up around identification of the best test selection procedure; the teacher estimation approach has grown up around the evaluation of different instructional programs. Both approaches can be conducted at any level of analysis thought appropriate (student, instructional group, class, or school; item, subtest, or total test).

Curriculum analysis approach. Approaches to curriculum analysis have tended to involve matches between detailed scope and sequence charts and test descriptions of content covered (Armbruster, Stevens, & Rosenshine, 1977; Everett, 1976; Kugle & Calkins, 1976; Pidgeon, 1970). These analyses are usually conducted at the total test and total curriculum level; they do not include information on how much material was actually covered in instruction by the school, class, or student. These analyses often look at both sides of the overlap question--how much of the test has been covered by the curriculum as well as how much of the curriculum was tested. In the Armbruster et al. (1977) analysis, the relationship between 3 curricula and 2 tests ranged from .10 to .43 using 6 out of 16 categories shared by all curricula and tests, but only a small percentage of skills taught were tested. In analyses of this type, it is assumed that the same or similar labels (e.g. detail, paraphrase, main idea, etc.) refer to the same content and that different labels refer to different content. In spite of rather gross measures of curriculum and test content, the approach clearly revealed several things. First, introductory reading curricula cover remarkably different content. (Beck and McCaslin, 1978, also showed this dramatically.) Second, tests (or at least those reviewed) cover a more similar range of topics than curricula. Third, very little of what is taught ever gets tested. Fourth, it is important to use such information in selecting a test for program evaluation.

Joseph Jenkins and Darlene Pany (1976) also noted the differences between content covered in reading curricula and the content of

standardized reading tests (or subtests). Their analysis of seven commercial reading series and five reading achievement tests was conducted by matching the words presented in each curriculum with the words that appeared on each test. Results revealed "curriculum bias" between tests for a single curriculum as well as on a single test for different reading curricula. In their discussion of the implications of these findings for educational research, they suggest that one must either control for curricula across treatment conditions or develop tests that are curriculum-based (or criterion-referenced).

More recently, Porter and his colleagues at the Institute for Research on Teaching have been analyzing tests and curricula in elementary mathematics (Floden, Porter, Schmidt, & Freeman, 1978; Kuhs, Schmidt, Porter, Floden, Freeman, & Schwille, 1979; Porter, 1978; Porter, Schmidt, Floden, & Freeman, 1978 a & b; Schmidt, 1978; Schwille, Porter, & Gant, 1979). In this impressive line of work, a detailed taxonomy of elementary mathematics topics was constructed which can be used to map out tests and curricula. As Porter et al. point out, district-level decision makers need to be keenly aware of the relationship between tests and curricula if they are to make sense of test results from different schools. What this line of work would ultimately lead to is policy discussion about appropriate content to be covered by instruction and subsequently to be tested. Thus, it could potentially inform decision-makers regarding curriculum design and selection and test selection. The main drawbacks of this approach for program evaluation are that it considers only material included in a formal curriculum, not teacher presentation, and does not yield a

metric that can be incorporated into an analysis.

A different approach to the overlap problem has emerged from program evaluation work in which the problems of multiple (in some cases, multi-national) curricula, a limited test battery and interpretation of results are prevalent. In an evaluation, if a particular test is used that reflects the content of one curriculum more than another, the test can be considered to be biased in favor of that curriculum. This seems obvious, but the precise measurement of how much more a test reflects the content of one curriculum than another is usually not undertaken nor is it usually included in the analysis of results. Just as it is inappropriate to examine posttest differences without pretest information, it is also inappropriate to ignore variation in the opportunity to learn the material that is being tested. Failure to consider variation in the opportunity to learn the material may lead to the attribution of differences to programs or student characteristics when, in fact, the differences may lie in the match (or mismatch) between what is being tested and what has been taught.

The earliest work in actually measuring overlap was by Husén (1967), Chang and Raths (1971), and Comber and Keeves (1973). In a fashion reminiscent of the scope and sequence charts, Comber and Keeves asked teachers to estimate the quantity of a test covered by instruction (teacher presentation plus curriculum) for an <u>entire school</u>. Husén apparently obtained slightly more detailed estimates by obtaining percentages of students that covered each test <u>item</u>, but the procedures were unevenly used across the study. These measures

focused only on how much of the test was covered by instruction, not the other way around. The measure obtained was too gross to adequately account for variation in student performance and was not included in the analyses, but was used as a guide for the interpretation of results.

The design of the Instructional Dimensions Study (IDS) included estimates of opportunity to learn content that was tested on a standardized test battery (Cooley & Leinhardt, 1975; 1980). This measure was to be included as a covariate in the analysis of the effectiveness of individualized instruction for compensatory programs. In implementing this aspect of the design, Lee Poynor (1978) used two basic estimation approaches, one based on teacher interviews at the end of the year, the other on a curriculum analysis of material covered by students. Both approaches are limited to estimating how much of the test was covered, not how much of what was taught was tested. In the next section, we turn to the specific details of measuring overlap directly.

Measuring Overlap

Teacher-Based Measures of Overlap

In IDS, the teacher estimate was obtained for every item of the test but at the class level (Poynor, 1978). Teachers were asked to estimate the percentage of students that had been taught the minimum material necessary to pass the item. Percentage of overlap scores were obtained by a conversion process that yields a somewhat ambiguous

number. The number reflects both the percentage of the test taught and the percentage of students that were taught it. Given that analyses were conducted at the classroom level, the information, while still less fine-grained than one might wish, was both usable and informative (Cooley & Leinhardt, 1980).

Another approach to obtaining a teacher estimate of overlap has been used in several instructional effectiveness studies (Cooley, Leinhardt, & Zigmond, 1979; Leinhardt & Engel, 1980; Leinhardt, Zigmond, & Cooley, 1980). In this approach, two components of a test item are considered, content and format. The teacher is asked to identify for each student (or a sample of students) whether or not the student has been taught the information required to answer the item, as well as whether or not the student has been exposed to the type of format the item employs. The teacher is not being asked whether s/he taught to the item, rather s/he is being asked whether the student has been taught the information the item is testing; for younger children, this includes familiarity with format. The teacher does this task for each student and each item; the time required to complete this task is minimal (approximately 30 minutes for 10 children). The teacher estimates include information about content covered through curriculum and teacher presentation, but may also include information about a teacher's expectation of success for a given student. The teacher estimates do not include information on how much of what was taught was tested.[1]

Curriculum-based Measures of Overlap

In order to avoid the problem of teacher bias another technique has been developed, computer-based curriculum analysis. This approach combines teacher sources of information with curriculum analysis. To use this approach, each item of a test is analyzed to assess what information is needed to "pass" the item. Information on the curricula used and the location (beginning and end) of each student in each curriculum is obtained. A dictionary of test-relevant information is then constructed (for example, vocabulary words presented in texts) for each student, and the dictionaries are matched with the item information to determine what percentage of the test has been covered through curriculum presentation. A similar system was used in IDS by Lee Poynor (1978). When multiple curricula are involved, even a small study involving 52 students (Cooley et al., 1979) can become very difficult. Vast amounts of information on each curriculum must be entered, sorted, and merged with student files. Further, while teachers can make some judgment about instructional adequacy for paragraph comprehension, computer-based curriculum analyses have so far been extremely conservative, and have drastically underestimated instructional coverage.

In summary, there are two basic ways in which estimates of instructional overlap with criterion measures have been obtained: teacher responses and computer-based analyses. Regardless of which approach is used, the information can be collected at the student, instructional unit, class, grade or school level; the entire test, or some smaller portion of it, may be the focus of the initial measure.

Thus, overlap can range from an estimate of what percentage of the total test has been covered by an entire school to which items have been covered by one student. Obviously, if data are collected at the student by item level, they can be aggregated to higher levels at the time of analysis.

Overlap in Evaluative Research

One of the important uses of overlap is in interpreting the results of an evaluation or of evaluative research. In this case, overlap is acting as a covariate in much the same way as pretest scores do. In a regression sense, the equation is:

$$Y = A + b_1 X_1 + b_2 X_2 + b_3 X_3 \ldots$$

Where Y is a posttest score, A is the intercept, X_1 is the pretest, X_2 is overlap, and X_3 ... is program membership or some cluster of process variables such as instructional time, teacher behaviors, etc. Figure 1 displays these variables in a causal map. Criterion performance is considered to be affected by what students knew initially, overlap, program membership and/or instructional processes; in addition, overlap is affected by what students knew initially and possibly by program membership. Of course, X_3, instructional process, may itself be more complexly modeled and, in some cases, causally linked to X_1. It should be noted that the general regression equation above does not test the model in Figure 1, it only tests the arrows impinging directly on posttest.

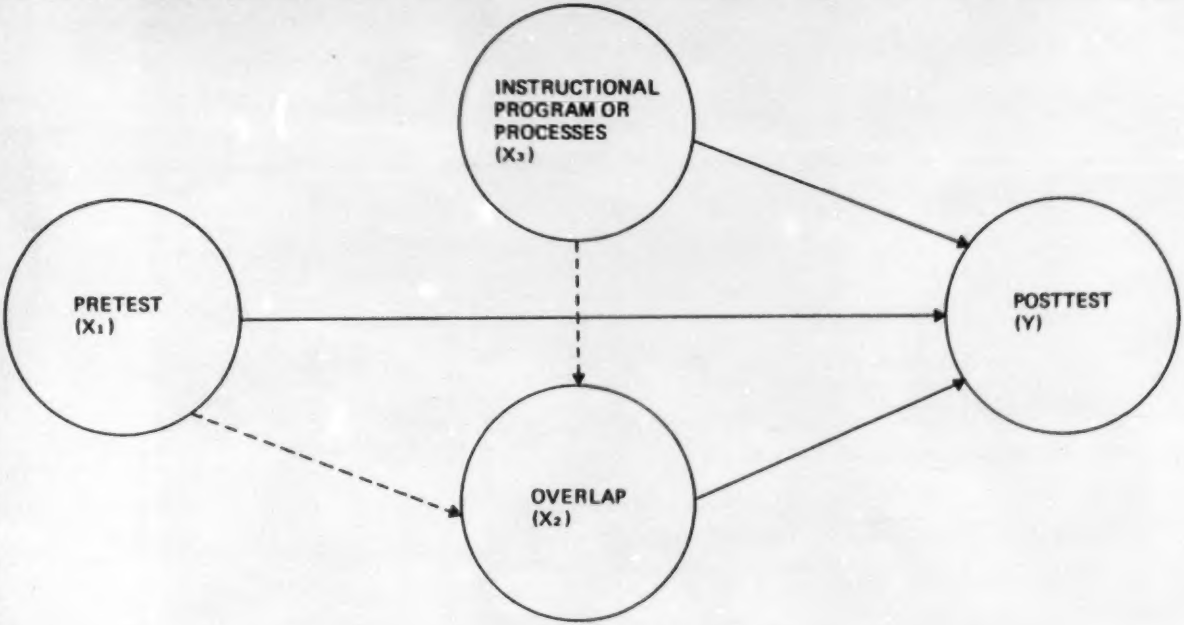

Figure 1. A Causal Map of Overlap in Evaluative Research

The importance of overlap for any given study is contingent upon many elements: the accuracy of estimation, the degree to which what has been taught has been learned, the degree to which what has not been taught has not been learned; and the complexity or non-hierarchical nature of the subject matter domain. The first three of these elements are somewhat self-explanatory, the fourth is less so. It is likely that as subject matter complexity increases and hierarchy decreases, overlap will be both less important and more difficult to measure with respect to a single instructional exposure such as a year. For example, a student's ability to write a cohesive argument about Macbeth as a tragic figure in Shakespeare, or to describe the relationship of membership in a cross-cousin matrilineal society and yam growing to ego development among the Trobriand Islanders is dependent not only on students having been exposed to Macbeth, Shakespearean drama, Freud, and Malinowski, but probably a vast number of other elements that permit students confronted with a new task or tasks to vary in their responses. However, elementary education in the basic skills (reading and arithmetic) is somewhat easier to analyze for purposes of estimating overlap.

Teacher-Based Estimates of Overlap

As previously mentioned, IDS included two estimates of overlap. In both cases, the pre and posttest of interest was the Comprehensive Tests of Basic Skills (CTBS) (CTB/McGraw-Hill, 1973). Approximately 400 first and third grade classrooms were studied during reading and mathematics instruction. The teacher's estimate of overlap was

obtained by asking teachers to determine the percentage of students who had been taught the information required by each item and then averaging across items to get a percentage of the test that had been covered. The means, standard deviations, correlations and regressions of the teachers' estimates of overlap are reported in Table 1.

As can be seen, before any program information has been included, pretest and overlap explain considerable and significant portions of the variances. Three of the means hover around 50 percent with 20 as a standard deviation; correlations with pretest are about .3 and with posttest about .4. The increase in R^2 from first to third grade is largely due to the stronger relationship between pre and posttest in the higher grades. This is reflected not only by the zero order correlations but also in the greater magnitude of the coefficients and the smaller standard errors. Also worth noting is how uncommon first grade math is, both in mean overlap estimates and in R^2. This, along with other information, represents a warning signal that first grade math results are of some concern.

Table 1

Analysis of Teachers' Estimates of Overlap (IDS)

	n	Overlap mean	Overlap s.d.	Correlations: Overlap with Pretest	Correlations: Overlap with Posttest	Correlations: Pretest with Posttest
Grade 1 Read (R 1)	104	50.93	18.46	.33	.47	.50
Grade 1 Math (M 1)	84	27.59	12.33	.32	.37	.39
Grade 3 Read (R 3)	109	51.12	21.33	.34	.38	.86
Grade 3 Math (M3)	116	56.14	19.40	.41	.51	.78

Regression Equations[a]	Adjusted R^2
$\text{Posttest}_{R1} = 30.9 + .63\,\text{Pretest} + .17\,\text{Overlap}$ (.14) (.04)	.34
$\text{Posttest}_{M1} = 20.1 + .52\,\text{Pretest} + .13\,\text{Overlap}$ (.19) (.05)	.20
$\text{Posttest}_{R3} = 15.8 + .89\,\text{Pretest} + .05\,\text{Overlap}$ (.06) (.02)	.74
$\text{Posttest}_{M3} = 16.0 + .91\,\text{Pretest} + .13\,\text{Overlap}$ (.08) (.03)	.66

[a] All of the coefficients and adjusted R^2 are significant at or below .05. Standard errors are in parentheses.

More recently, we have been involved in a study of reading in elementary level Learning Disabilities classrooms. This study was begun during the 1977-1978 school year and continued through 1978-1979 (Cooley, W. W., Leinhardt, G., & Zigmond, N., 1979; Leinhardt, G., Zigmond, N., & Cooley, W. W., 1980). In the first year, the CTBS was used as both the pre and posttest measure; during the 1978-1979 phase, the pretest used was the Spache Diagnostic Reading Scales (Spache, 1972). Overlap data were collected at the student by item level. Teachers were asked, for each student, to circle items (on the reading subtests of the CTBS) that contained content covered by instruction whether the instruction was in text or classroom teaching. Overlap was the number of items circled divided by the total number of possible items, times 100. The estimates of 52 cases from the first year of the study ranged from 2.70 to 100.00 percent ($\overline{X}$ = 59.12; s.d. = 32.73). In the second year of the study, the estimates ranged from 7.14 to 100 percent for 105 cases ($\overline{X}$ = 56.33; s.d.= 27.33).

Table 2

Correlations and Regressions

Using Teachers' Estimates of Overlap (LD)

Correlation Matrix (1977-1978)

		1	2	3	4
1.	Pretest	1.00	.73	.56	.82
2.	Teacher's Estimate of Overlap	.73	1.00	.45	.81
3.	Silent Reading (per 40 minutes)	.56	.45	1.00	.62
4.	Posttest	.82	.81	.62	1.00

Posttest[a] = 134.1 + .40 Pretest + .64 Overlap + 34.4 Silent Reading
(.10) (.14) (12.8)

Adjusted R^2 = .79

Correlation Matrix (1978-1979)

		1	2	3	4
1.	Pretest	1.00	.40	.63	.83
2.	Teacher's Estimate of Overlap	.40	1.00	.31	.50
3.	Silent Reading (per day)	.63	.31	1.00	.63
4.	Posttest	.83	.50	.63	1.00

Posttest[a] = 177.6 + 6.2 Pretest + .40 Overlap + 1.1 Silent Reading
(.66) (.12) (.45)

Adjusted R^2 = .72

[a]All of the coefficients and adjusted R^2 are significant at or below .05. Standard errors are in parentheses.

Table 2 presents the correlations and regressions for the two years of study. Overlap was obtained by teacher interviews for each child and is the percentage of the test covered by instruction. Silent reading is a measure of the amount of time a student spends reading silently, and was obtained by observing individual students during regular classroom instruction. This variable is included to represent one of the most relevant aspects of instructional processes as depicted in Figure 1. Overlap is again an important variable in predicting end-of-year test performance. In this set of evaluative research studies, the important aspect is not which program the student was following, but the relationship between student behaviors and reading performance. Here again, knowledge of the degree of overlap is critical to interpreting the results.[2]

Curriculum-based Estimates of Overlap

Curriculum-based estimates of overlap are much harder to obtain than teacher estimates, but, on the surface at least, they have more objectivity and are less subject to bias. Two studies have used a curriculum-based estimate, IDS (Cooley & Leinhardt, 1980; Poynor, 1978) and the first year of the LD reading study (Cooley et al., 1979).

Lee Poynor developed a curriculum-based estimate of overlap at the student level for IDS. The curriculum-based estimate used teacher reports of content covered in each text for each student and matched that content to the content of the test. The results are presented in Table 3.

Table 3

Analysis of Curriculum-based Estimates of Overlap (IDS)

		Overlap		Correlations		
	n	mean	s.d.	Overlap with Pretest	Overlap with Posttest	Pretest with Posttest
Grade 1 Read (R 1)	104	27.13	14.91	.21	.42	.50
Grade 1 Math (M 1)	84	15.03	10.38	.33	.38	.39
Grade 3 Read (R 3)	109	20.02	5.60	-.05	.10	.86
Grade 3 Math (M 3)	116	30.52	14.56	.30	.42	.78

Regression Equations[a]

		Adjusted R^2
$Posttest_{R1}$	= 31.95 + .70 Pretest (.13) + .20 Overlap (.05)	.34
$Posttest_{M1}$	= 21.4 + .52 Pretest (.19) + .16 Overlap (.06)	.20
$Posttest_{R3}$	= 11.8 + .94 Pretest (.05) + .25 Overlap (.08)	.75
$Posttest_{M3}$	= 17.2 + .95 Pretest (.08) + .15 Overlap (.04)	.65

[a]All of the coefficients and adjusted R^2 are significant at or below .05. Standard errors are in parentheses.

Three interesting differences between Table 1 and Table 3 should be pointed out. First, all estimates are lower using curriculum analyses rather than teacher estimates. Second, Grade 1 math, while lowest, is not as dramatically different from the rest of the curriculum-based estimates as it is in the teacher estimates. Third, Grade 3 reading overlap does not correlate with either pretest or posttest, however, the regression coefficient is still significant. This is probably due to the difficulty in estimating content covered prior to grade 3 in reading. In order to estimate item-level overlap, some assumptions must be made about what the student was exposed to prior to Book 3 Level 1, for example. The regression results are almost identical to those obtained using the teacher estimate. Considering the substantial differences in means and the totally different process of gathering the information, this suggests that estimates are somewhat stable regardless of technique.

A curriculum-based measure of overlap was also included in the design of the first year of the study of reading in learning disabilities classrooms (1977-1978). In January of 1978, teachers were asked to list the major curricula used with each student. At the time of posttesting, May, 1978, that list was verified and teachers were asked where each student was in each curriculum at that point in time (i.e., final location). For each level of each curricular series, all words presented were entered into the computer with an identifier indicating the series, level, and unit, chapter, or page in which the word appeared. These words were then sorted in alphabetical order and duplications were deleted based on the higher level

identifiers. These words then formed a dictionary of unique words including the first presentation of each word only. Separate dictionaries were compiled for each curriculum. Individual student dictionaries were compiled to include those levels completed in each curriculum the student used during the year based on the student's end-of-year location given by the teacher. These were matched with the appropriate level of the posttest (CTBS).

Once all dictionaries had been entered, verified, sorted, duplicates deleted and reverified, a computer program was designed by Melanie Bowen to match individual student dictionaries with the CTBS dictionaries. The program then calculated the percent of overlap in several ways. The total test by item analysis (as opposed to a by word analysis ignoring items) is reported here. The total test by item measure of curricular overlap had a mean of 19.65 (s.d. = 14.73; n = 52).

The results shown in Table 4 again indicate that while the means are lower for curriculum-based estimates, the regression is essentially the same. These results, coupled with the IDS results, suggest that the mean curriculum-based overlap estimate is always lower than the teacher estimate because it automatically leaves out in-class instruction not found in textbooks, but that both estimates do equally well in predicting posttest. Choosing which estimate is better is a matter of either philosophy or money.

Table 4

Correlations and Regressions

Using Curriculum Estimates of Overlap

		1	2	3	4
1.	Pretest	1.00	.67	.56	.82
2.	Curriculum Estimate of Overlap	.67	1.00	.59	.71
3.	Silent Reading (per 40 minutes)	.56	.59	1.00	.62
4.	Posttest	.82	.71	.62	1.00

Posttest[a] = 107.6 + .59 Pretest + .70 Overlap + 28.2 Silent Reading
(.11) (.36) (15.6)

Adjusted R^2 = .72

---

[a] All of the coefficients and adjusted R^2 are significant at or below .05.

Standard errors are in parentheses.

Summary and Conclusions

In studying the effectiveness of different instructional practices or programs, it is possible that an outcome measure may be biased in favor of a particular practice or program because the overlap between the test and one program is greater than for the other(s). When this occurs, one program or set of practices may look artifically better with respect to test performance than another. Awareness of this problem is long standing. Effective ways of dealing with the problem are just emerging.

In addition to test modification or construction, two basic approaches for dealing with the overlap issue have emerged. The first approach is a systematic analysis of curricula and tests to help guide test selection. The most promising work for primary level mathematics has been done by Porter and his colleagues. Information from this analysis (if the analyses are expanded) can serve as a basis for critiquing tests and curricula, and can aid policy analysts in interpreting research results. A second approach to overlap is to measure the degree of overlap directly, and incorporate such a measure into the analysis. This approach could be used in conjunction with the first, or alone.

Two ways to directly measure overlap have been developed. The first involves teacher interviews or questionnaires. The second involves analyzing the curriculum to assess if information required by the test has been covered by the curriculum. The teacher interview approach reflects both informal in-class instruction and curriculum-based instruction, but it may also include the teacher's

expectation about student competency. [It is worth noting that in the study of reading instruction discussed earlier, an estimate of teacher expectation for academic success failed to predict teachers' estimates of overlap. Thus, overlap estimates seem to be freer of teacher bias than we originally assumed (Leinhardt et al., 1980).] The curriculum analysis approach is less likely to be biased, but it is costly, time-consuming, and less likely to capture informal instruction. Both approaches do equally well in predicting final test performance.

In addition to the two measurement approaches discussed, some attention needs to be paid to the level at which the data are collected. It is our conviction that student by item level data are the easiest and the most accurate. If time or cost preclude gathering the information for each student on each item, then students should be randomly sampled and data aggregated. Having the teacher estimate percentages of students requires the teacher to think of groups of students, estimate the percentages they represent, and average (2 out of 30 have all, 3 out of 30 have none, etc.). The task is quite complex and likely to be errorful.

In conclusion, in order to assure that evaluation results do not misrepresent programmatic or instructional differences, it is vital to include information about overlap in the analysis. Future work should lead to the improvement of measurement techniques (perhaps including frequency of presentation of information) and to a greater understanding of what types and levels of criterion tasks will not be predicted by simple estimates of overlap.

References

Armbruster, B. B., Stevens, R. J., & Rosenshine, B. Analyzing content coverage and emphasis: A study of three curricula and two tests (Technical Report 26). Urbana, IL: University of Illinois at Urbana-Champaign, 1977.

Beck, I. & McCaslin, E. An analysis of dimensions that affect the development of code-breaking ability in eight beginning reading programs. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center, 1978. (LRDC Publication No. 1978/6)

Chang, S. S. & Raths, J. The school's contribution to the cumulating deficit. Journal of Educational Research, 1971, 64, 272-276.

Cole, N. S. & Nitko, A. J. Instrumentation and bias: Issues in selecting measures for educational evaluation. Paper presented at the National Symposium on Educational Research, Johns Hopkins University, November 1979.

Comber, L. C. & Keeves, J. P. Science education in nineteen countries. International Studies in Evaluation I. New York: John Wiley & Sons, 1973.

Cooley, W. W. & Leinhardt, G. Design for the Individualized Instruction Study: A study of the effectiveness of individualized instruction in the teaching of reading and mathematics in compensatory education programs. Final Report. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center, 1975.

Cooley, W. W. & Leinhardt, G. The Instructional Dimensions Study. Educational Evaluation and Policy Anaylsis, 1980, 2(1), 7-25.

Cooley, W. W., Leinhardt, G., & Zigmond, N. Explaining reading performance of learning disabled students. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center, 1979. (LRDC Publication No. 1979/12)

CTB/McGraw-Hill. Comprehensive Tests of Basic Skills. Monterey, CA: McGraw-Hill, 1973.

Everett, B. E. A preliminary study of the relevance of a standardized test for measuring achievement gains in innovative arithmetic programs. Project Longstep Final Report: Volume II, Appendix Report. Palo Alto, CA: American Institutes for Research, 1976.

Fisher, C. W., Berliner, D. C., Filby, N. N., Marliave, R., Cahen, L. S., Dishaw, M. M., & Moore, J. E. Teaching and learning in the elementary school: A summary of the Beginning Teacher Evaluation Study. Beginning Teacher Evaluation Study, Technical Report VII-1. San Francisco, CA: Far West Laboratory for Educational Research and Development, 1978.

Floden, R. E., Porter, A. C., Schmidt, W. H., & Freeman, D. J. Don't they all measure the same thing? Consequences of selecting standardized tests. East Lansing, MI: Institute for Research on Teaching, Michigan State University, July, 1978. (Research Series No. 25)

Husén, T. (Ed.) International study of achievement in mathematics: A comparison of twelve countries. Volume II. New York: John Wiley & Sons, 1967.

Jenkins, J. R. & Pany, D. Curriculum biases in reading achievement tests. Technical Report No. 16, Center for the Study of Reading, University of Illinois at Urbana-Champaign, November, 1976.

Kugle, C. L. & Calkins, D. S. The effect of considering student opportunity to learn in teacher behavior research (Research Report No. 7). Austin, TX: University of Texas, Research and Development Center for Teacher Education, 1976.

Kuhs, T., Schmidt, W., Porter, A., Floden, R., Freeman, D., & Schwille, J. A taxonomy for classifying elementary school mathematics content. East Lansing, MI: Institute for Research on Teaching, Michigan State University, April, 1979. (Research Series No. 4)

Leinhardt, G. & Engel, M. Iterative evaluation: NRS, an example. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Leinhardt, G., Zigmond, N., & Cooley, W. W. Reading instruction and its effects. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Marliave, R., Fisher, C., Filby, N. & Dishaw, M. The development of instrumentation for a field study of teaching. Beginning Teacher Evaluation Study, Technical Report I-5. San Francisco, CA: Far West Laboratory for Educational Research and Development, 1977.

Pidgeon, D. A. Expectation and pupil performance. Stockholm: Almquist & Wiksell, 1970.

Popham, W. J. The case for criterion-referenced measurements. Educational Researcher, 1978, 7(11), 6-10.

Porter, A. C. Relationships between testing and the curriculum. East Lansing, MI: Institute for Research on Teaching, Michigan State University, July, 1978. (Occasional Paper No. 9)

Porter, A. C., Schmidt, W. H., Floden, R. E., & Freeman, D. J. Impact on what?: The importance of content covered. East Lansing, MI: Michigan State University, Institute for Research on Teaching, 1978. (Research Series No. 2) (a)

Porter, A. C., Schmidt, W. H., Floden, R. E., & Freeman, D. J. Practical significance in program evaluation. American Educational Research Journal, 1978, 15(4), 529-539. (b)

Poynor, L. Instructional Dimensions Study: Data management procedures as exemplified by curriculum analysis. Paper presented at the annual meeting of the American Educational Research Association, Toronto, April 1978.

Rosenshine, B. Academic engaged minutes, content covered, and direct instruction. Unpublished manuscript, University of Illinois at Urbana-Champaign, 1978.

Schmidt, W. H. Measuring the content of instruction. East Lansing, MI: Institute for Research on Teaching, Michigan State University, October, 1978. (Research Series No. 35)

Schwille, J., Porter, A., & Gant, M. Content decision-making and the politics of education. East Lansing, MI: Michigan State University, Institute for Research on Teaching, June, 1979. (Research Series No. 52)

Spache, G. D. Diagnostic reading scales. Monterey, CA: CTB/McGraw-Hill, 1972.

Walker, D. F. & Schaffarzick, J. Comparing curricula. Review of Educational Research, 1974, 44, 83-112.

Footnotes

1. Using a related measure of teacher diagnostic skill, the researchers in the BTES study scored the teacher responses in terms of their accuracy (Fisher, Berliner, Filby, Marliave, Cahen, Dishaw, & Moore, 1978; Marliave, Fisher, Filby, & Dishaw, 1977). The distinction is important. To obtain an overlap measure, one merely sums the items the teacher estimates have been covered by instruction. To obtain an estimate of diagnostic skill (or hits), one sums the number of items for which the teacher's estimate and student performance concur. For example, if a teacher said none of the material on a test had been taught, the overlap measure would be zero; if the students missed all of the items on such a test, the diagnostic score would be 100 percent.

2. A second measure calculated was the percent of hits, or diagnostic ability. A hit was counted for each item for which the teacher's estimate matched the actual performance of the student. The percent of hits for 1977-1978 ranged from 47.30 to 94.59 percent with a mean of 70.74 percent (s.d.=12.22). The correlation with posttest was .53 and .58 with pretest. In 1978-79 with 105 cases, the range was 43.24 to 91.89 percent with a mean of 64.30 percent (s.d.=10.72), and the correlation with the posttest is .20 and .14 with pretest.